Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers

Okoniewski, Michal J ; Meienberg, Janine ; Patrignani, Andrea ; Szabelska, Alicja ; Matyas, Gabor ; Schlapbach, Ralph

Abstract: Herein we present the applicability of single-molecule (PacBio RS) and second-generation sequencing technology (Illumina) to the characterization of large genomic deletions. By testing samples previously characterized using a Sanger approach, our methods determined that both next-generation sequencing platforms were able to identify the position of deletion breakpoints. Our results point out various advantages of next-generation sequencing platforms when characterizing genomic deletions; however, special attention must be dedicated to identical sequences flanking the breakpoints, such as poly(N) motifs.

DOI: https://doi.org/10.2144/000113992

Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers

Michal J Okoniewski [1,2*†], Janine Meienberg [3,4*], Andrea Patrignani [1], Alicja Szabelska [1,5], Gabor Matyas [3,4#], Ralph Schlapbach [1#]

[1] Functional Genomics Center Zurich, Zurich, Switzerland

[2] Department of Neuroimmunology and Multiple Sclerosis Research, Neurology Clinic, University Hospital, Zurich, Switzerland

[3] Center for Cardiovascular Genetics and Gene Diagnostics, Zurich, Switzerland

[4] Zurich Center of Integrative Human Physiology, University of Zurich, Zurich, Switzerland

[5] Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poznan, Poland

* These authors contributed equally to this work

# These authors jointly directed this work

† Correspondence: Michal J Okoniewski, michal.okoniewski@fgcz.ethz.ch, FGCZ, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Keywords: sequencing, large deletions, amplicons, PacBio, Illumina, Sanger

Word count for Abstract: 83.

Word count for the whole manuscript: 2520, for manuscript body: 1118.

Abstract: *This study presents the applicability of single molecule (PacBio RS) and second generation sequencing technology (Illumina) to the characterization of large genomic deletions. Methods have been tested on samples previously characterized using a Sanger approach. By the means of both next generation sequencing platforms, we were able to identify the position of deletion breakpoints. The obtained results point out various advantages of next generation sequencing platforms, while special attention must be dedicated to identical sequences flanking the breakpoints, such as a poly(N) motif.*

The PacBio technology has not only the potential to identify modified bases and thus to characterize methylation patterns [1,2] but it also provides previously unprecedented sequencing read lengths (>2kb) and is hence useful for quick improvement of existing genome assemblies [3]. In this study, we used the advantage of such long reads for the characterization of large deletions previously identified by multiplex ligation-dependent probe amplification (MLPA) and microarray analyses. Otherwise, using traditional Sanger sequencing the characterization of large deletions is time consuming and work intensive [4,5], increasing the need for effective breakpoint localization. Indeed, for Sanger sequencing a large fragment (2-10kb) containing the breakpoints has to be amplified by long-range PCR (LR-PCR) and subsequently sequenced in order to be able to identify the exact position of breakpoints. As by means of Sanger sequencing just ~600bp can be sequenced using one primer, it needs several sets of internal primers for a large LR-PCR product.

In contrast, next generation sequencing (NGS) may offer simplified sequencing in such cases. Here, we tested this possibility by using not only long reads of the PacBio platform but also short reads of a second generation sequencing technology (Illumina). Illumina offers stable length of short reads (100bp in this case) with errors most likely to be grouped at the ends of reads [6,7], while in our hands PacBio reads had a mean length of 2459bp and random distribution of errors affecting 10-15% of nucleotides. In addition, only few dedicated computational techniques are available for the characterization of large deletions by NGS [8], making data analysis a challenge.

The three DNA samples used in this study carry previously characterized large hemizygous deletions, two of which with a size of 26,887bp (sample 44) and 302,580bp (sample 70), respectively, affect the *FBN1* gene in patients with Marfan syndrome [4] and one of which with a size of 3,408,306bp comprises the entire *COL3A1* gene in a patient with Ehlers-Danlos syndrome vascular type (sample 53B) [5]. Accordingly, ~6.5-8.5kb LR-PCR products were amplified using the Expand Long Template PCR System (Roche Diagnostics, Rotkreuz, Switzerland) as described previously [4,5] and purified by means of QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).

SMRTbell libraries were prepared using the PacBio C2 chemistry (3-10kb) DNA preparation kit (Part# 001-540-726, Pacific Biosciences, Menlo Park, CA, USA) as well as 5µg purified amplicons without fragmentation. Libraries were subsequently sequenced on the PacBio RS (Pacific Biosciences) using one SMRT cell per sample and taking two movies of 45 minutes each. The reads have been mapped with the BLASR mapper [9], which is supplied in the SMRT Portal software suite (Pacific Biosciences) and applies therefore as standard mapper for PacBio reads. The same amplicons have been

sequenced on a HiSeq 2000 sequencer (Illumina, San Diego, CA, USA) using Illumina's TruSeq DNA Sample Preparation v2 protocol with 1µg input material and 100+100bp pair-end reads. The reads have been mapped using the standard mapper bowtie [10]. For both NGS platforms the mappers have been used with default parameters and respective sequences are available in the SRA archive (study ID: ERP002092).

For PacBio data, the read coverage in the SMRT Portal software suite resulted in a clear drop of read depth in the deleted region (see Supplementary Figure 5), which was subsequently confirmed by zooming in on the breakpoint regions by means of the Integrative Genomics Viewer (IGV) [11] (Figure 1 and Tables 1-2 for sample 70, supplementary Figures 1 and 2 as well as supplementary Tables 1-4 for samples 44 and 53B, respectively). Respective Illumina data displayed in IGV show much more gradually sinking patterns at the expected deletion ends and the site of breakpoints in these data was identified by an increase in mismatches (Figure 1, Tables 1-2, supplementary Figures 1 and 2, supplementary Tables 1-4). This may be expected by the fact that the mappers typically allow several mismatches and thus many of the short Illumina reads could be mapped over the breakpoints. In contrast, in case of PacBio there is a number of reads spanning over the deletion, which have not been mapped by the SMRT Portal aligner to the standard reference due to the high number of mismatches. The read depth of both platforms is more than sufficient to find the breakpoint – tests with half or one third of the data gave also satisfactory results (data not shown).

An additional difficulty may be identical sequences on both sides of the deletion, a common phenomenon that has already been described for different genes [12-14]. In particular, this could be observed in all three deletions presented in this study ("CC" in

samples 53B and 70 as well as “GC” in sample 44). In order to find the precise sequence at the sites of break and rejoining with poly(N) motifs (tandemly repeated nucleotides), we have developed an AWK script to count matches at the sites of suspected deletion breakpoints (s. script in supplementary data). This counting was performed with perfect matches only, resulting in the data depicted in Figure 2 (sample 70) and supplementary Figures 3 and 4 (samples 44 and 53B, respectively). When a single nucleotide (or pair in the case of GC) has a fixed probability of being misinterpreted, it can be assumed without loss of generality that the distribution of the occurrences of specific motifs follows the Poisson distribution. The hypothesis that the maximum of counts represents the appropriate motif has been tested. At significance level fixed to 0.01, the probabilities of wrongly accepting null hypothesis are for PacBio reads in sample 70 equal to 1.5e-23, 1.06e-46, and 3.2e-141 in the cases of 20, 10, and 5 flanking bases, respectively (Figure 2). In the case of Illumina, due to the high number of reads, the error levels are so low that they go below the small number precision in the R language. For details on the calculations, see the R script in supplementary data. The script can be used on any fasta or fastq data and checks the statistical power at a given significance level regardless of the platform.

As shown by this study, the determination of deletion breakpoints can be done with data obtained from both NGS platforms. However, whereas the long reads of PacBio RS showed a sharp decrease in read depth, in short Illumina reads it was rather an increase in mismatches related to the position of the breakpoints. Sample preparation costs are comparable for PacBio and Illumina. However, sequencing using PacBio can be done within a working day, while Illumina even in the smaller MiSeq version requires more time. In conclusion both platforms are suitable for precise breakpoint localization and

hence provide an alternative procedure for the characterization of large deletions, which is much less resource and time consuming than traditional Sanger sequencing.

Supplementary data

1. Supplementary Figures 1-5
2. Supplementary Tables 1-4
3. Example of an AWK script for counting exact matches in Figure 2
4. R script for calculation of corresponding type II errors
5. Sequence data deposited in the SRA archive (ERP002092)

Acknowledgements

We are grateful to Yu-Chih Tsai, Jonas Korlach, and Stephen W. Turner for discussion on PacBio technology and data analysis. This work was supported by the FGCZ as well as grants from the COFRA Foundation (to GM), Gottfried & Julia Bangerter-Rhyner-Stiftung (to GM), Jubiläumsstiftung Swiss Life (to GM), Foundation for People with Rare Diseases (to JM and GM), Clinical Research Priority Program (CRPP/KFSP-MS) of University of Zurich (to MO), and Sciex.ch (nr. 11.182 to AS and MO).

Competing interests

The authors claim no competing interests.

References

1. **Clark, T.A., I.A. Murray, R.D. Morgan, A.O. Kislyuk, K.E. Spittle, M. Boitano, A. Fomenkov, R.J. Roberts, and J. Korlach.** 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res *40*:e29.

2. **Murray, I.A., T.A. Clark, R.D. Morgan, M. Boitano, B.P. Anton, K. Luong, A. Fomenkov, S.W. Turner, et al.** 2012. The methylomes of six bacteria. Nucleic Acids Res *40*:11450-11462.

3. **Zhang, X., K.W. Davenport, W. Gu, H.E. Daligault, A.C. Munk, H. Tashima, K. Reitenga, L.D. Green, and C.S. Han.** 2012. Improving genome assemblies by sequencing PCR products with PacBio. Biotechniques *53*:61-62.

4. **Matyas, G., S. Alonso, A. Patrignani, M. Marti, E. Arnold, I. Magyar, C. Henggeler, T. Carrel, et al.** 2007. Large genomic fibrillin-1 (FBN1) gene deletions provide evidence for true haploinsufficiency in Marfan syndrome. Hum Genet *122*:23-32.

5. **Meienberg, J., M. Rohrbach, S. Neuenschwander, K. Spanaus, C. Giunta, S. Alonso, E. Arnold, C. Henggeler, et al.** 2010. Hemizygous deletion of COL3A1, COL5A2, and MSTN causes a complex phenotype with aortic dissection: a lesson for and from true haploinsufficiency. Eur J Hum Genet *18*:1315-1321.

6. **Kozarewa, I., Z. Ning, M.A. Quail, M.J. Sanders, M. Berriman, and D.J. Turner.** 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods *6*:291-295.

7. **McElroy, K.E., F. Luciani, and T. Thomas.** 2012. GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics *13*:74.

8. **Ye, K., M.H. Schulz, Q. Long, R. Apweiler, and Z. Ning.** 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*:2865-2871.

9. **Chaisson, M.J. and G. Tesler.** 2012. Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. BMC Bioinformatics *13*:238.

10. **Langmead, B.** 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics *Chapter 11*:Unit 11.7.

11. **Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov.** 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform [Epub ahead of print].

12. **Giacalone, J.P. and U. Francke.** 1992. Common sequence motifs at the rearrangement sites of a constitutional X/autosome translocation and associated deletion. Am J Hum Genet *50*:725-741.

13. **Otto, E., R. Betz, C. Rensing, S. Schatzle, T. Kuntzen, T. Vetsi, A. Imm, and F. Hildebrandt.** 2000. A deletion distinct from the classical homologous recombination of juvenile nephronophthisis type 1 (NPH1) allows exact molecular definition of deletion breakpoints. Hum Mutat *16*:211-223.

14. **Liu, H.X., L. Cartegni, M.Q. Zhang, and A.R. Krainer.** 2001. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat Genet *27*:55-58.

Figure legends

Figure 1. Both ends of the 302,580-bp deletion on chromosome 15 in sample 70 displayed in the Integrative Genomics Viewer (aligned reads are displayed as grey bars/arrows, letters indicate mismatched bases, purple vertical dashes insertions, and black horizontal lines deletions). The reads generated by PacBio RS (upper panel) as well as Illumina HiSeq 2000 (lower panel) were sorted by aligned position, base, and mapping quality and compared to the results of Sanger sequencing (bottom). Sections of 22 reads are shown. The top tracks show much clearer decrease in the PacBio reads, whereas the Illumina reads show a clearer increase in mismatches at the sites of breakpoints. Uppercase letters represent the sequence in the region of the start point of the deletion and lowercase letters the sequence in the region of the deletion end point. Due to identical sequences at the site of breakpoints, the break and re-joining could have occurred at three positions as indicated by open triangles. The dotted red line marks the most telomeric position of the possible breakpoints. Note that the total read counts (reads) and the percentage of reference bases (%) are given for the positions flanking the site where the coverage (grey bars) starts to lower (for more details see Tables 1 and 2).

Figure 2. Counts of exact matches for different lengths of a poly(C) motif (red) at the site of deletion breakpoint in sample 70 with 5-, 10-, and 20-nucleotide flanking sequences for both PacBio and Illumina reads (denoted by different colors), indicating that the true sequence includes 4×C (n=4) (cf. Figure 1). Corresponding type II errors were calculated using the R script provided in supplementary data.

Tables

Table 1. Read depth and percentage of wild-type allele in the region flanking the breakpoint at the start site of the deletion in sample 70.

Location	Not deleted					Deleted				
Wild-type sequence	A	C	C	**C**	**C**	C	C	A	T	T
PacBio RS	614 (100%)	531 (100%)	508 (100%)	492 (100%)	384 (99%)	103 (89%)	66 (88%)	66 (98%)	61 (95%)	29* (86%)
HiSeq 2000	12699 (100%)	11679 (100%)	9545 (100%)	8278 (100%)	7081 (100%)	5820 (1%)	4313 (48%)	3224 (100%)	1678 (99%)	1536** (98%)

* No mapped reads 243 bases after the most telomeric breakpoint (read depth = 0).

** No mapped reads 117 bases after the most telomeric breakpoint (read depth = 0).

Bold letters indicate identical bases at the site of breakpoint, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the three possible breakpoints. This is also the point where the read depth drops and the number of mismatches increases (cf. Figure 1).

Table 2. Read depth and percentage of wild-type allele in the region flanking the breakpoint at the end of the deletion in sample 70.

Location	Deleted					Not deleted				
Wild-type sequence	t	t	t	a	a	**c**	**c**	a	t	a
PacBio RS	231* (8%)	242 (10%)	251 (94%)	215 (95%)	315 (52%)	1316 (99%)	1359 (99%)	1411 (99%)	1465 (99%)	1547 (100%)
HiSeq 2000	5408** (100%)	8762 (87%)	10427 (83%)	13679 (92%)	14956 (62%)	16383 (100%)	17735 (96%)	18658 (100%)	19512 (100%)	20549 (100%)

* No mapped reads 210 bases before the most telomeric breakpoint (read depth = 0).

** No mapped reads 16 bases before the most telomeric breakpoint (read depth = 0).

Bold letters indicate identical bases at the site of breakpoints, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the possible breakpoints. The most centromeric breakpoint, where the read depth drops and the number of mismatches increases, is indicated by a black bold line (cf. Figure 1).

Figure 1

Click here to download Figures (separate file for each figure): Figure_1.pdf

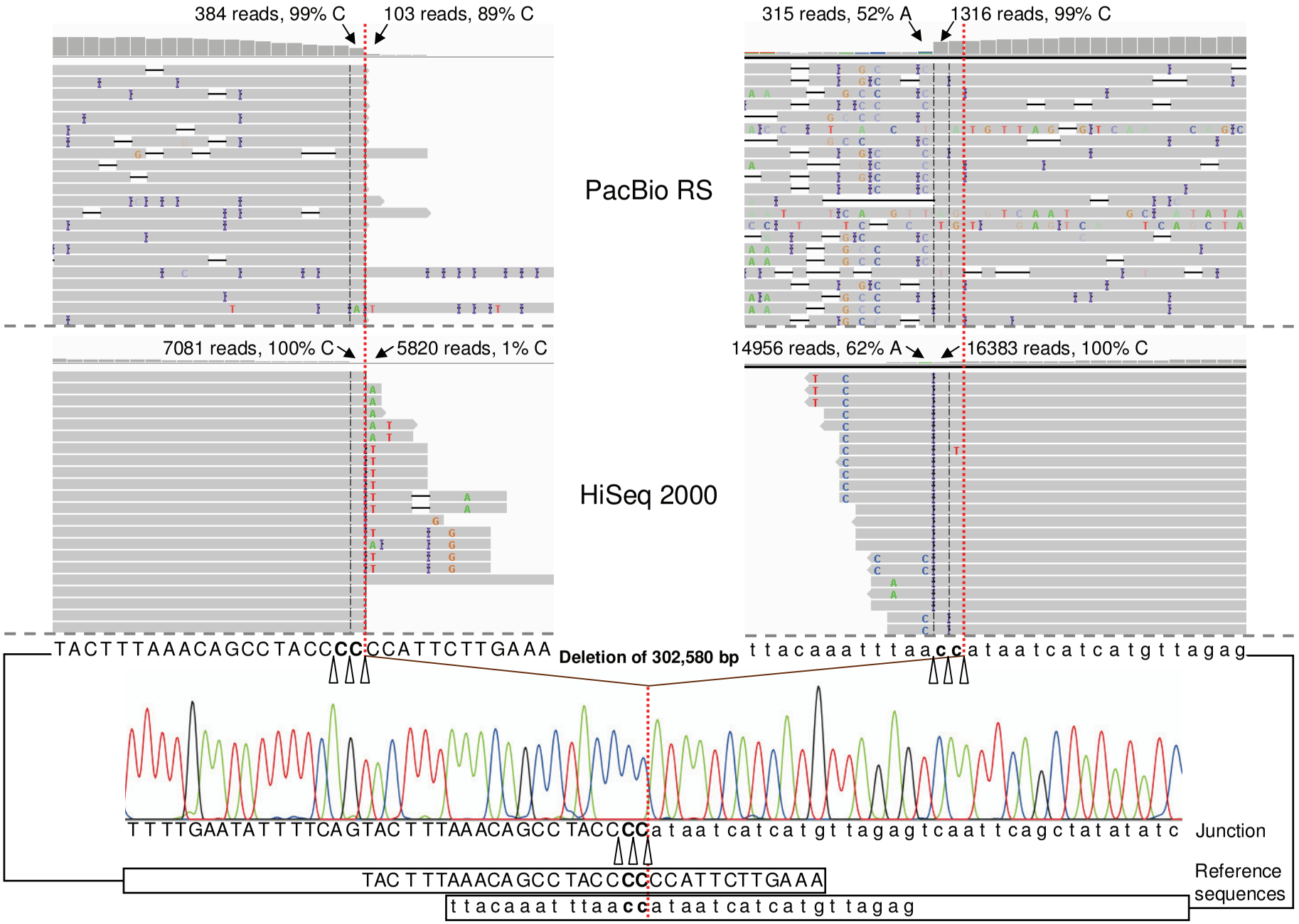


Figure 2
Click here to download high resolution image

PacBio										
	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	type II error
Flank 5	56	28	191	941	2132	348	37	8	0	3e-141
Flank 10	0	1	53	329	820	124	12	3	0	1e-46
Flank 20	0	0	5	56	127	19	2	1	0	5e-06
Illumina										
	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	type II error
Flank 5	2	3	0	29	68817	3	0	0	0	<2e-308
Flank 10	0	0	0	20	58769	0	0	0	0	<2e-308
Flank 20	0	0	0	13	43258	0	0	0	0	<2e-308

Supplementary awk script

Click here to download Supplementary Material (supporting information for online publication only): check.awk.txt

```
BEGIN {c8=0; c7=0; c6=0; ;c5=0; c4=0; c3=0; c2=0; c1=0; c0=0; sekw="C"; l=0}
{
 if ($0 ~ />/)
  {
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCCCATAATCATCATGTTAGAGTC/) c8=c8+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCCATAATCATCATGTTAGAGTC/) c7=c7+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCCATAATCATCATGTTAGAGTC/) c6=c6+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCCATAATCATCATGTTAGAGTC/) c5=c5+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCCCATAATCATCATGTTAGAGTC/) c4=c4+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCCATAATCATCATGTTAGAGTC/) c3=c3+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACCATAATCATCATGTTAGAGTC/) c2=c2+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTACATAATCATCATGTTAGAGTC/) c1=c1+1;
   if (sekw ~ /TCAGTACTTTAAACAGCCTAATAATCATCATGTTAGAGTC/) c0=c0+1;
   l=l+length(sekw)
   sekw = "";
  }
 else sekw=sekw$0;
}
END { print (c0," ",c1," ",c2," ",c3," ",c4," ",c5," ",c6," ",c7," ", c8," ",l)}
```

Supplementary R script

Click here to download Supplementary Material (supporting information for online publication only): typeIIerror_estimate.R.txt

```
dane<-read.csv('polyCmotifs.csv', header=T) #loading data with counts of the exact matches with
5,10,20-nucleotide flanking region
a<-0.01 # significance level of the test

z<-qnorm(1-a) # z statistics needed for the calculations of the type II error
n<-rowSums(dane) # total number of the number of
l1<-40
l1<-c(10,20,40) #lenght of flanking region
p<-0.12 # probability of mismatch
N<-83830 # total number of reads
psum1<-dnbinom(l1,1,p) #estimated probability that there is no error in the sequence, negative
binomial distribution is assumed
n.reads1<-N*psum1 # expected number of exact matches with chosen flanking region


#calculation of the type II error (probability that the null hypothesis was wrongly accepted)
beta<-NULL
x<-seq(0,8,by=1) #considered number of deletions

for (i in 1:3)
{
 lambda<-as.numeric(colnames(dane)[which(dane[i,]==max(dane[i,]))]) # choice of the motif with
maximun counts of the exact matches
 x<-x[-lambda] #obtaining possible alternative hypotheses by exculing the lambda from considered
cases

 #calculations of errors has to be divided in 2 cases, when alternative is smaller or higher than null
hypothesis
 y<-x[x<lambda]
 b<-1-pnorm((lambda-z*sqrt(lambda/n[i])-y)/sqrt(y/n[i])) # type II error for alternative hypotheses
< lambda

 y<-x[x>lambda]
 b<-c(b,pnorm((lambda+z*sqrt(lambda/n[i])-y)/sqrt(y/n[i])))# type II error for alternative
hypotheses > lambda

 beta<-c(beta,sum(b))
}

print(beta)
```

Supplementary Figure 1. Both ends of the 26,887-bp deletion on chromosome 15 in sample 44 displayed in the Integrative Genomics Viewer. The reads generated by PacBio RS (upper panel) as well as Illumina HiSeq 2000 (lower panel) were sorted by aligned position, base, and mapping quality and compared to the results of Sanger sequencing (bottom). Sections of 22 reads are shown. Symbols and labels are as used in Figure 1. For more details see supplementary Tables 1 and 2.

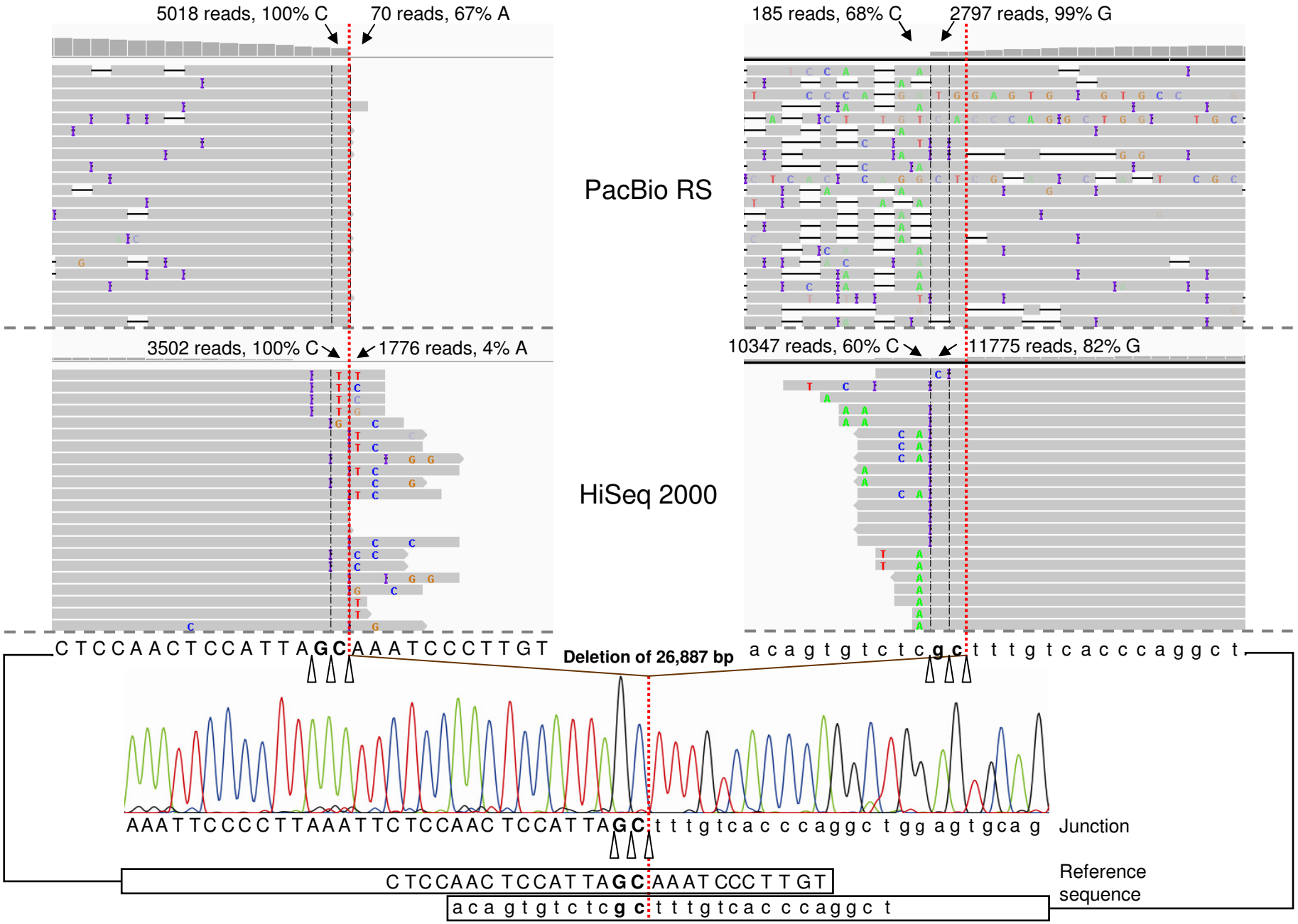
Supplementary Figure 2. Both ends of the 3,408,306-bp deletion on chromosome 2 in sample 53B displayed in the Integrative Genomics Viewer. The reads generated by PacBio RS (upper panel) as well as Illumina HiSeq 2000 (lower panel) were sorted by aligned position, base, and mapping quality and compared to the results of Sanger sequencing (bottom). Sections of 22 reads are shown. Symbols and labels are as used in Figure 1. For more details see supplementary Tables 3 and 4.

Supplementary Figure 3. Counts of exact matches for different lengths of a GC-motif (red) at the site of deletion breakpoints in sample 44 with 5-, 10-, and 20-nucleotide flanking sequences for both PacBio and Illumina reads (denoted by different shades of green), indicating that the true sequence includes 1×GC (n=1) (cf. supplementary Figure 1). Corresponding type II errors were calculated using the R script provided in supplementary data.

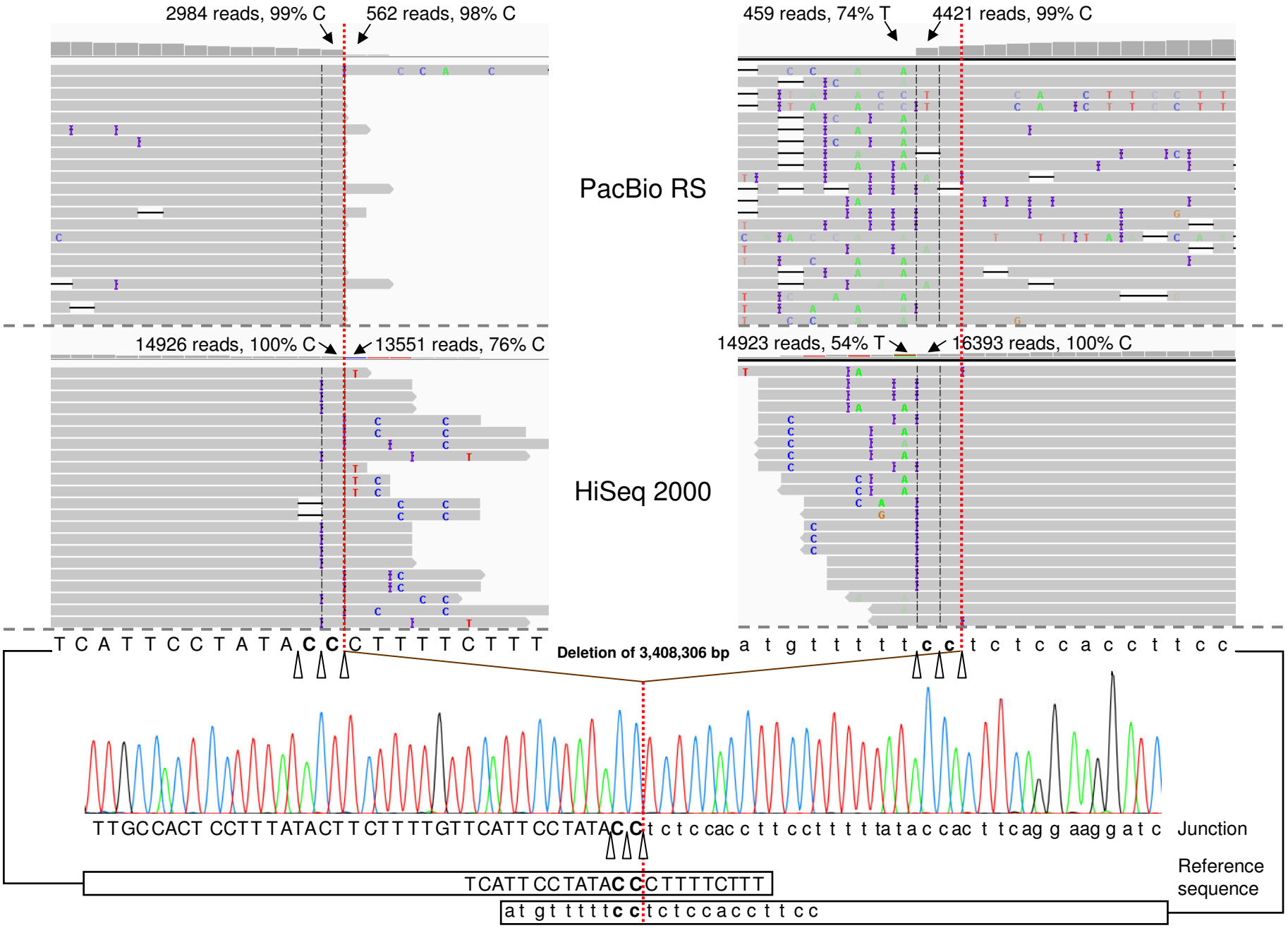
Supplementary Figure 4. Counts of exact matches for different lengths of a poly(C) motif (red) at the site of deletion breakpoints in sample 53B with 5-, 10-, and 20-nucleotide flanking sequences for both PacBio and Illumina reads (denoted by different shades of green), indicating that the true sequence includes 2×C (n=2) (cf. supplementary Figure 2). Corresponding type II errors were calculated using the R script provided in supplementary data.

Supplementary Figure 5. Coverage plot from the PacBio software – SMRTportal, based upon a standard human genome reference.

supplementary Figure 1



supplementary Figure 2



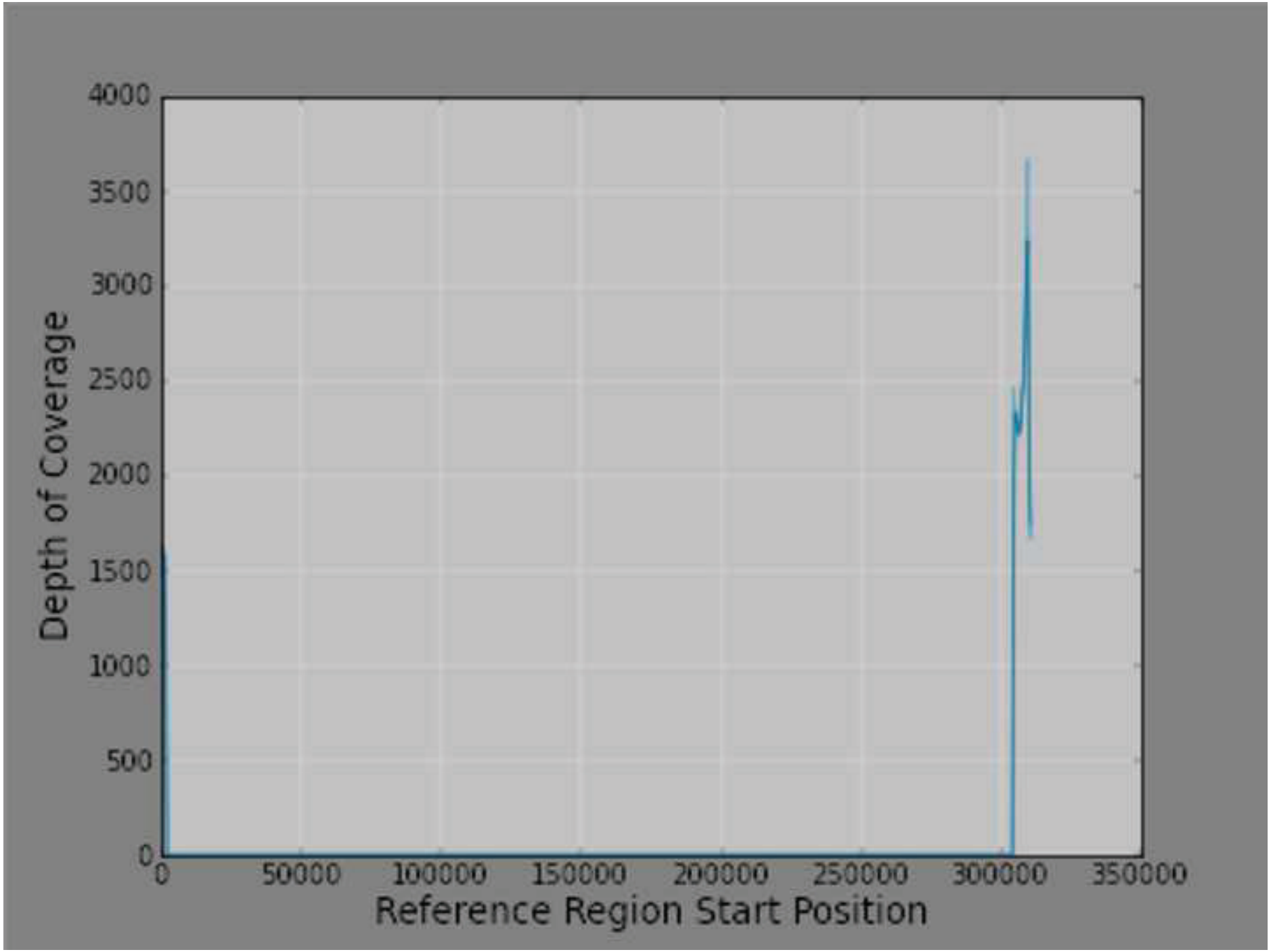
Supplementary Figure 3
Click here to download high resolution image

PacBio				
	n=0	n=1	n=2	type II error
Flank 5	885	2270	11	<2e-308
Flank 10	0	649	4	6e-61
Flank 20	0	74	1	4e-06
Illumina				
	n=0	n=1	n=2	type II error
Flank 5	116	52218	0	<2e-308
Flank 10	0	44249	0	<2e-308
Flank 20	0	32717	0	<2e-308

Supplementary Figure 4
Click here to download high resolution image

PacBio						
	n=0	n=1	n=2	n=3	n=4	type II error
Flank 5	200	888	2904	244	12	6e-280
Flank 10	19	253	1026	61	2	3e-84
Flank 20	3	41	186	12	1	6e-13
Illumina						
	n=0	n=1	n=2	n=3	n=4	type II error
Flank 5	14	1	53212	0	0	<2e-308
Flank 10	0	0	45448	0	0	<2e-308
Flank 20	0	0	32554	0	0	<2e-308

Supplementary Figure 5
Click here to download high resolution image



Supplementary Tables

Click here to download Supplementary Material (supporting information for online publication only): Supplementary_Tables.doc

Supplementary Table 1: Read depth and percentage of wild-type allele in the region flanking the breakpoint at the start site of the deletion in sample 44.

Location	Not deleted					Deleted				
Wild-type sequence	T	T	A	**G**	**C**	A	A	A	T	C
PacBio RS	7226 (99%)	6738 (99%)	6144 (99%)	5621 (99%)	5018 (100%)	70 (67%)	35 (40%)	43 (37%)	150 (96%)	39* (79%)
HiSeq 2000	9403 (100%)	7894 (100%)	6118 (100%)	4626 (100%)	3502 (100%)	1776 (4%)	845 (2%)	92 (20%)	15 (53%)	6** (67%)

* No mapped reads 468 bases after the most telomeric breakpoint (read depth = 0).
** No mapped reads 14 bases after the most telomeric breakpoint (read depth = 0).
Bold letters indicate identical bases at the site of the breakpoint, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the three possible breakpoints. This is also the point where the read depth drops and the number of mismatches increases (cf. Supplementary figure 1).

Supplementary Table 2: Read depth and percentage of wild-type allele in the region flanking the breakpoint at the end of the deletion in sample 44.

Location	Deleted					Not deleted				
Wild-type sequence	g	t	c	t	c	**g**	**c**	t	t	t
PacBio RS	77* (34%)	104 (72%)	71 (76%)	132 (72%)	185 (68%)	2797 (99%)	3061 (99%)	3495 (99%)	4025 (99%)	4219 (100%)
HiSeq 2000	2200** (4%)	5594 (69%)	8918 (87%)	9609 (83%)	10347 (60%)	11775 (82%)	13756 (100%)	14409 (100%)	14934 (99%)	15995 (100%)

* No mapped reads 368 bases before the most telomeric breakpoint (read depth = 0).
** No mapped reads 39 bases before the most telomeric breakpoint (read depth = 0).
Bold letters indicate identical bases at the site of the breakpoints, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the possible breakpoints. The most centromeric breakpoint, where the read depth drops and the number of mismatches increases, is indicated by a black bold line (cf. Supplementary figure 1).

Supplementary Table 3: Read depth and percentage of wild-type allele in the region flanking the breakpoint at the start site of the deletion in sample 53B.

Location	Not deleted					Deleted				
Wild-type sequence	A	T	A	**C**	**C**	C	T	T	T	T
PacBio RS	4386 (99%)	4132 (99%)	3931 (99%)	3568 (99%)	2984 (99%)	562 (98%)	493 (96%)	211 (60%)	190 (59%)	192* (92%)
HiSeq 2000	20179 (100%)	18897 (100%)	17702 (100%)	14635 (100%)	14926 (100%)	13551 (76%)	12489 (70%)	11602 (71%)	8538 (99%)	8470** (60%)

* No mapped reads 544 bases after the most telomeric breakpoint (read depth = 0).
** No mapped reads 13 bases after the most telomeric breakpoint (read depth = 0).
Bold letters indicate identical bases at the site of the breakpoint, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the three possible breakpoints. This is also the point where the read depth drops and the number of mismatches increases (cf. Supplementary figure 2).

Supplementary Table 4: Read depth and percentage of wild-type allele in the region flanking the breakpoint at the end of the deletion in sample 53B.

Location	Deleted					Not deleted				
Wild-type sequence	t	t	t	t	t	**c**	**c**	t	c	t
PacBio RS	223* (81%)	230 (79%)	240 (61%)	331 (91%)	459 (74%)	4421 (99%)	5192 (100%)	5678 (100%)	6066 (100%)	6516 (100%)
HiSeq 2000	7818** (73%)	9888 (100%)	10675 (57%)	13335 (89%)	14923 (54%)	16393 (100%)	17832 (100%)	19429 (100%)	20646 (100%)	21893 (100%)

* No mapped reads 411 bases before the most telomeric breakpoint (read depth = 0).
** No mapped reads 11 bases before the most telomeric breakpoint (read depth = 0).
Bold letters indicate identical bases at the site of the breakpoints, which can be either up- or downstream of the breakpoint. The red dotted line indicates the most telomeric position of the possible breakpoints. The most centromeric breakpoint, where the read depth drops and the number of mismatches increases, is indicated by a black bold line (cf. Supplementary figure 2).